Stock Picking Resized by Market Regime Risk

A survival game for mean-variance investing

Giovanni Sella (gs956) Marcus Chung (mlc264) Matteo Cozzani (mc3954) Wendy Zhao (yz2586)

MGT 924 Statistical Foundations · Problem Set 3 · JKP Common Task Framework · October 2026

1. Introduction

Every month an investor in US stocks has to answer two separate questions: which stocks to hold, and how much risk to take with them. We answer each with a model built on its own strand of the asset-pricing literature. The choice of stocks rests on a factor model of the 402 characteristics, built on the evidence of Jensen, Kelly and Pedersen (2023) that published return anomalies are real and organize into a small number of themes; it delivers a market-neutral long-short book. The amount of risk is set by a hidden Markov model of the stock market, in the tradition of Hamilton (1989), which estimates whether the market is calm or turbulent. The estimated regime serves as a forecast of how risky the book is about to be, and the book is sized accordingly, following the logic of volatility-managed portfolios (Moreira and Muir 2017).

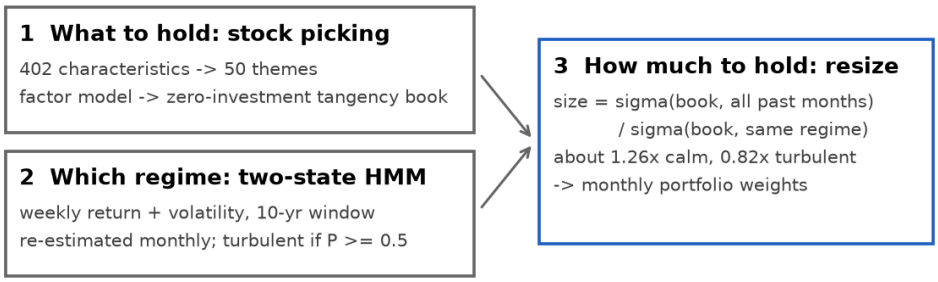


Figure 1: At each month-end the theme model gives the stocks to hold, the HMM gives the market regime, and the sizing rule sets how much of the book to hold.

2. Stock selection: a survival game on characteristic portfolios

Jensen, Kelly and Pedersen (2023, JKP) show that published return characteristics cluster into a small number of themes whose members measure a common premium. We follow this structure with two changes: themes are formed by co-failure, and members are weighted by their survival in random replays of history.

A characteristic is a number attached to each stock, so its return has to be constructed. Following the standard construction of characteristic-sorted factors, we turn it into a trade: each month, stocks are ranked on the characteristic, those ranked high are bought and those ranked low are sold short, with positions proportional to the rank so that long and short positions offset. The next-month return of this long-short portfolio is the characteristic's return; its history since 1952 yields an average return, the premium, and a standard deviation. Each portfolio is signed so that its past average is positive. Within the game, monthly returns are standardized by the portfolio's own standard deviation, because a 3% loss is unremarkable for momentum, whose monthly volatility is close to 4%, but extreme for an accounting signal with 1% volatility. A standardized value of -2 therefore denotes a loss twice the usual size for that portfolio.

Both the grouping of characteristics and their weights within groups are determined by a survival game played on realized history. A trial draws past months at random and with replacement, restricted to months whose returns were known at the time, so that each trial replays history with real months in a random order and combination. Table 1 summarizes the three steps of the game.

Step	Method: what happens in the game	Rationale: principle and academic counterpart
1. Group (co-failure)	*10,000 months are drawn. In each, a signal fails if it ranks in the worst third of all signals. The φ coefficient (the correlation of two signals' failure indicators) measures how much more often a pair fails together than by chance; hierarchical clustering with Ward linkage on the distance $\sqrt{2(1-\varphi)}$ forms the themes. All 402 signals are kept, each assigned to one theme.*	*Signals that fail together are the same bet. Co-failure replaces the return correlation used by JKP; joint failure measures downside (lower-tail) dependence, analogous to the coexceedances of Bae, Karolyi and Stulz (2003).*
2. Weight within themes (survival)	*A trial draws 12 months from each third of the available history (early, middle, recent). A signal loses the trial if in any of the 36 months its standardized return is below -2, and wins otherwise; any number of signals can win a trial. Within a theme, weights are proportional to the survival rate over 1,000 trials.*	*The survival rate measures the reliability of the premium in every era, and has a Spearman rank correlation of 0.45 with the pre-1990 premium. A return below -2σ is a Value-at-Risk exceedance (probability 2.3% under normality), so the survival rate is the frequency of zero exceedances in 36 months ($0.977^{36} \approx 0.44$ for a normal signal), the event counted by a VaR backtest (Kupiec 1995). It replaces JKP's equal weights; expected returns enter only through the themes' average premia.*
3. Number of themes	*We run 2-fold cross-validation on 1952–1989, assigning alternate months to the two folds. For each candidate number of themes, from 5 to 100 and all 402 signals unclustered, the themes and the portfolio are estimated on one fold and the Sharpe ratio is computed on the other, and then the reverse. Fifty themes give the highest validation Sharpe ratio, and every smaller number is more than one standard error below it, so we keep 50.*	*Occam's razor: among choices that perform about equally well we keep the simplest, the smallest number of themes within one standard error of the best (one-standard-error rule). The standard error of a Sharpe ratio under i.i.d. returns follows Lo (2002). Only pre-1990 data are used, so the choice is fixed before the test period.*

Table 1: The survival game. It decides what the bets are (step 1) and which characteristics represent them (step 2); the variance-covariance matrix decides how much of each bet to hold.

Steady characteristics survive most trials. In the pre-1990 data, short-term reversal survives 83% of trials and value 73%, whereas momentum survives only 36%, because its occasional crashes eliminate it. Within the momentum theme, the steadier members consequently receive larger weights than the crash-prone ones. The game therefore rewards the reliability of a premium, and since survival is required in each of the three eras, characteristics that paid in only one period are penalized (Figure 2).

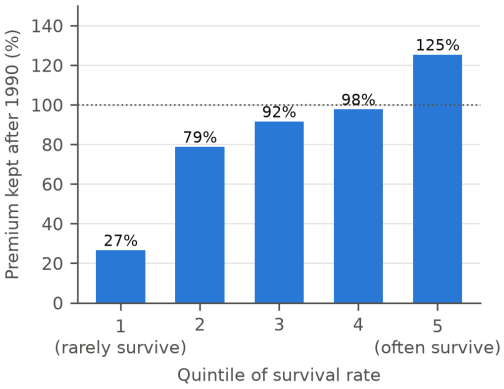


Figure 2: Median ratio of the post-1990 to the pre-1990 premium of the 402 characteristics, by quintile of survival rate. Survival is measured on 1952–2022, so the figure describes the characteristics, and Section 5 gives the trading evidence.

Given the themes, we estimate a Barra-style factor model with cross-sectional regressions. A stock's exposure (beta) to a theme is the survival-weighted average of its signed characteristic ranks, standardized each month. Each month t an OLS regression across stocks,

$$r_{i,t+1} = f_{0,t+1} + \beta'_{i,t} f_{t+1} + u_{i,t+1},$$

gives the theme returns f_{t+1} and the residuals $u_{i,t+1}$. This is the cross-sectional step of Fama and MacBeth (1973); since the exposures are observed characteristics, no first-step time-series regression for betas is needed, and the theme premia are the time-series averages of f_t, as in the second Fama–MacBeth step.

The mean-variance portfolio is $w \propto \Sigma^{-1}\mu$, but with thousands of stocks the sample mean and variance-covariance matrix are so noisy that the estimated optimum is dominated by estimation error. We therefore impose the factor structure:

$$\hat{\Sigma} = \beta\, \hat{\Sigma}_f\, \beta' + \mathrm{diag}(\hat{\sigma}_u^2), \qquad \hat{\mu} = \beta\, \hat{\mu}_f,$$

where β is the matrix of theme betas (one row per stock, one column per theme), $\hat{\Sigma}_f$ is the variance-covariance matrix of the theme returns shrunk toward a scaled identity (Ledoit and Wolf 2004), $\hat{\sigma}_u^2$ are the stocks' residual variances over the last 36 months, shrunk halfway toward the cross-sectional median, and $\hat{\mu}_f$ are the theme premia. The intercept's premium is set to zero and the weights sum to zero, so the book is market-neutral; it is scaled to unit gross exposure.

3. Market regimes

Market volatility clusters: it alternates between calm and turbulent regimes and is persistent enough to be forecast. We capture it with a two-state regime-switching model (Hamilton 1989).

From daily returns we construct a value-weighted market index and divide each month into weekly blocks ending on the month-end; each block yields a pair of observations, the return and the log realized volatility, so that the model sees both how much the market moved and how violently. The model has two hidden states, calm and turbulent, and within each state the pair follows a mixture of two bivariate Gaussians, one for ordinary weeks and one for shock weeks, which captures fat tails and the negative relation between returns and volatility (24 parameters). Parameters are estimated by the EM (Baum–Welch) algorithm from eight seeded starting points, and the state with the higher average volatility is labelled turbulent. The forward filter gives the probability of each state using data up to month-end t only, and this filtered probability is the one used by the model:

$$P(S_t = s \mid \text{data up to } t) \;\propto\; p(x_t \mid s) \sum_{s'} A_{s's}\, P(S_{t-1} = s' \mid \text{data up to } t-1).$$

The model is re-estimated every month-end on the last ten years of weekly observations. This specification, two Gaussians per state and a ten-year window, was selected among six candidates by out-of-sample predictive log-likelihood on 1972–1989, retaining the fewest Gaussians within one standard error of the best.

4. Position sizing under regime risk

For a mean-variance investor the optimal position in a strategy is proportional to its expected return divided by its variance, both conditional on what is known today:

$$w_t \propto \frac{E_t[r_{t+1}]}{\mathrm{Var}_t[r_{t+1}]}.$$

When the regime predicts variance while the mean stays constant, the position should shrink as risk rises. Scaling by inverse variance raises Sharpe ratios (Moreira and Muir 2017), and long-short books are most exposed to crashes in turbulent markets (Daniel and Moskowitz 2016). The size of the book at month-end t is

$$\mathrm{size}_t = \left(\frac{\sigma(\text{book, all past months})}{\sigma(\text{book, past months in the regime of } t)} \right)^p .$$

Both volatilities use only returns realized before t and require at least 24 months of history. Four sizing rules are candidates: the book alone ($p = 0$), volatility equalization ($p = 1$), the mean-variance rule with a constant mean ($p = 2$), and a switch to cash in turbulent months, which is optimal only if the book's expected return in those months is zero or negative. The rule with the highest Sharpe ratio before 1990 is selected: $p = 1$ (5.21, against 5.20 for $p = 2$, 5.05 for $p = 0$, 3.01 for switching to cash). The book is then held at about 1.26× in calm months and 0.82× in turbulent months (Figure 3): it is held throughout, at a smaller size when the market is risky.

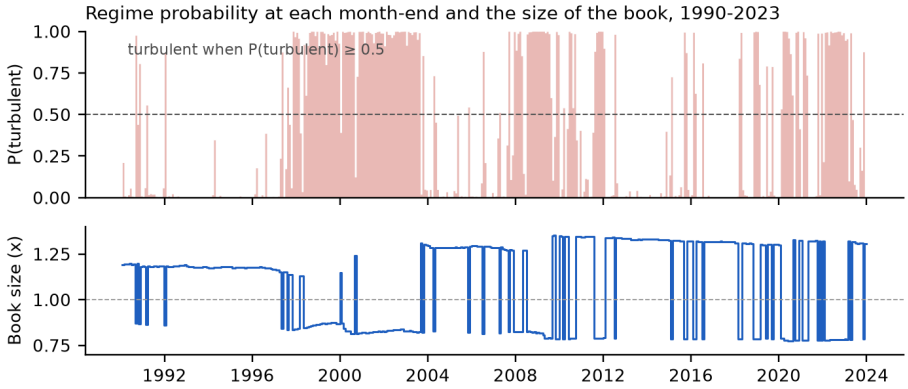


Figure 3: Top: probability that the market is turbulent at each month-end (forward filter, monthly re-estimation). Bottom: the resulting size of the stock-selection book.

5. Results, 1990–2023 (out of sample, before costs)

1990–2023, out of sample	Sharpe	Max drawdown (10% vol)	Corr. with market
Stock picking resized by regime	2.63	−16%	−0.06
Stock picking alone	2.46	−19%	−0.07
Market, value-weighted	0.54	−37%	1.00

Over 1990–2023 the stock-picking book earns a Sharpe ratio of 2.46, close to the organisers' Markowitz-ML benchmark (2.52), with almost no correlation with the market. Sizing by regime raises the Sharpe ratio to 2.63 (95% block-bootstrap interval for the gain: 0.06 to 0.28) and reduces the maximum drawdown from

−19% to −16%.

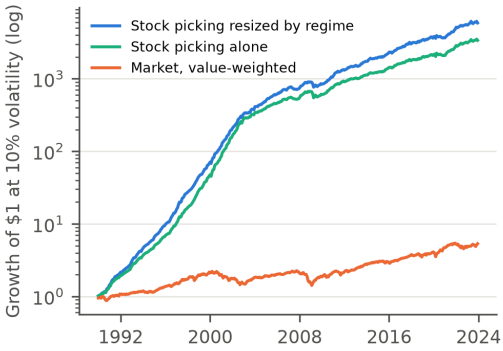


Figure 4: Growth of $1 in each strategy scaled to 10% annual volatility, 1990–2023 (log scale).

6. Disclosures

All design choices were fixed on pre-1990 data. Alternatives tested on that period (grouping by return correlation, other survival rules, discarding rarely surviving themes, weighting themes by survival in place of the covariance matrix, and switching the book to cash in turbulent months) all performed below the final design. Returns are excess returns, Sharpe ratios are annualized from monthly returns, and drawdowns are computed at 10% volatility, as on the leaderboard. AI tools were used to help write and test the code and this document; the authors are responsible for both.

References

Bae, K.-H., G. A. Karolyi, and R. M. Stulz (2003). A new approach to measuring financial contagion. *Review of Financial Studies* 16(3), 717–763.

Daniel, K., and T. J. Moskowitz (2016). Momentum crashes. *Journal of Financial Economics* 122(2), 221–247.

Fama, E. F., and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy* 81(3), 607–636.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *Review of Financial Studies* 33(5), 2223–2273.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57(2), 357–384.

Jensen, T. I., B. Kelly, and L. H. Pedersen (2023). Is there a replication crisis in finance? *Journal of Finance* 78(5), 2465–2518.

Jensen, T. I., B. Kelly, S. Malamud, and L. H. Pedersen (forthcoming). Machine learning and the implementable efficient frontier. *Review of Financial Studies*.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives* 3(2), 73–84.

Ledoit, O., and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88(2), 365–411.

Lo, A. W. (2002). The statistics of Sharpe ratios. *Financial Analysts Journal* 58(4), 36–52.

Moreira, A., and T. Muir (2017). Volatility-managed portfolios. *Journal of Finance* 72(4), 1611–1644.